Run Charts

Summary

The **Run Chart** procedure plots data contained in a single numeric column. It is assumed that the data are sequential in nature, consisting either of individuals (one measurement taken at each time period) or subgroups (groups of measurements at each time period). Tests are performed on the data to determine whether they represent a random series, or whether there is evidence of mixing, clustering, oscillation, or trending.

Sample StatFolio: *runchart.sgp*

Sample Data:

The file *bottles.sgd* contains the measured bursting strength of *n* = 100 glass bottles, similar to a dataset contained in Montgomery (2005). Each row consists of a sample tested at 10 minute intervals. The table below shows a partial list of the data from that file:

strength	*time*
255	0:10
232	0:20
282	0:30
260	0:40
255	0:50
233	1:00
240	1:10
255	1:20
254	1:30
259	1:40
235	1:50
262	2:00

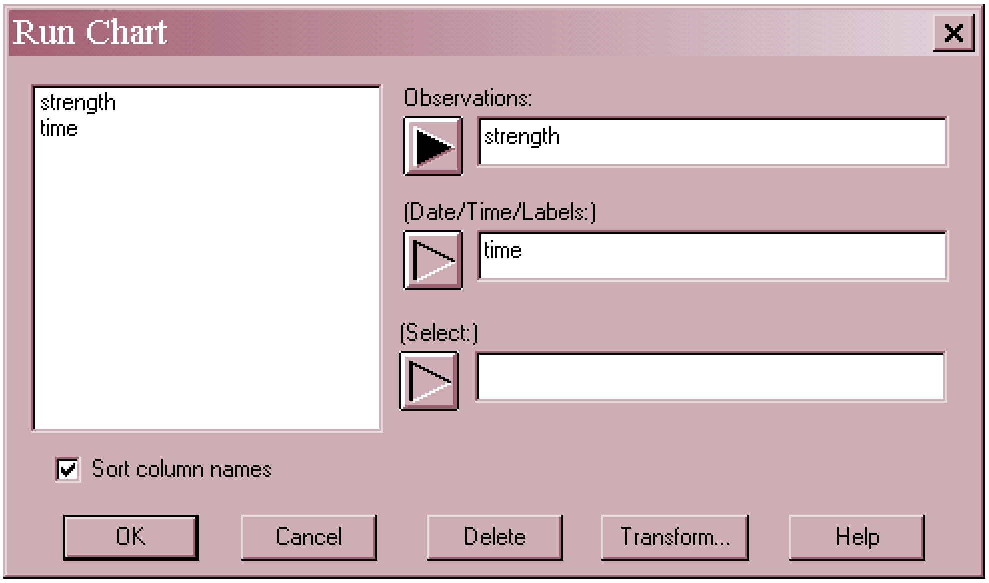
Strength is measured in pounds per square inch (psi).

Data Input

There are two menu selections that create a runs chart, one for individuals data and one for grouped data.

Case #1: Individuals

The data to be analyzed consist of a single numeric column containing *n* = 2 or more observations. The data are assumed to have been taken one at a time, in sequential order by rows.



Observations: strength

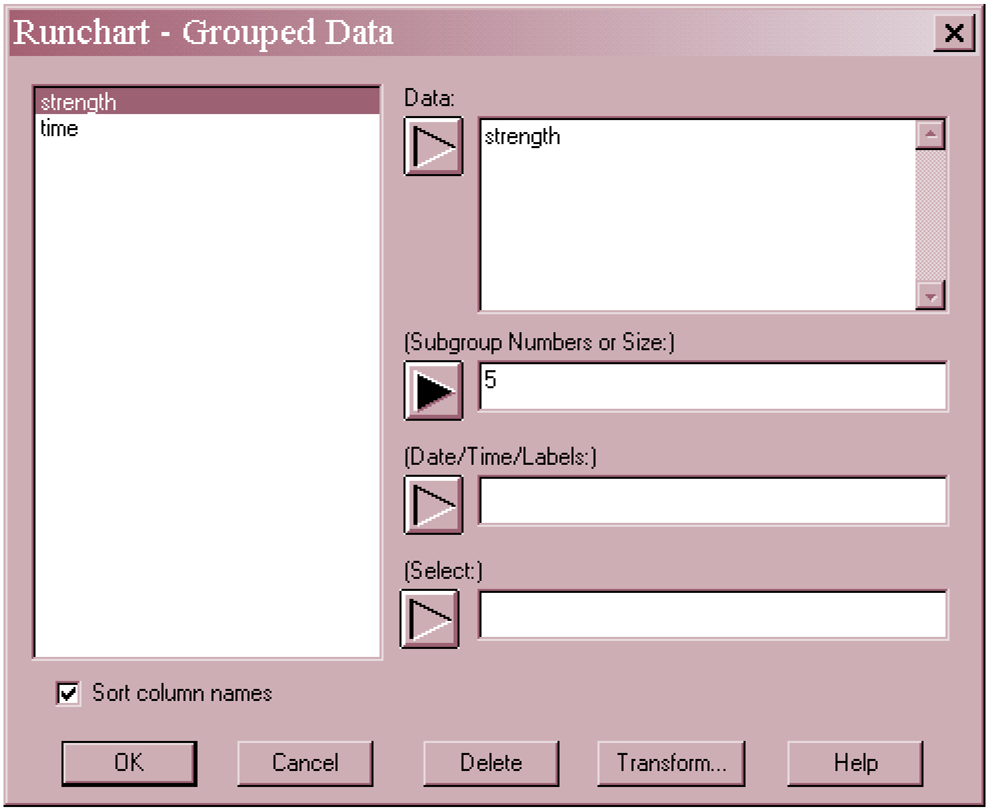
(Date/Time/Labels:) time

(Select:)

- **Observations:** numeric column containing the data to be analyzed.
- **Date/Time/Labels:** optional column containing row identifiers, used to scale the X axis. If this field is left blank, then row numbers are used to identify observations.
- **Select:** subset selection.

Case #2: Grouped Data

The data to be analyzed consist of one or more numeric columns. The data are assumed to have been taken in groups, in sequential order by rows.



Runchart - Grouped Data

strength
time

Data: strength

(Subgroup Numbers or Size:) 5

(Date/Time/Labels:)

(Select:)

☑ Sort column names

- **Data:** one or more numeric columns. If more than one column is entered, each row of the file is assumed to represent a subgroup with subgroup size *m* equal to the number of columns entered. If only one column is entered, then the *Subgroup Numbers or Size* field is used to form the groups.

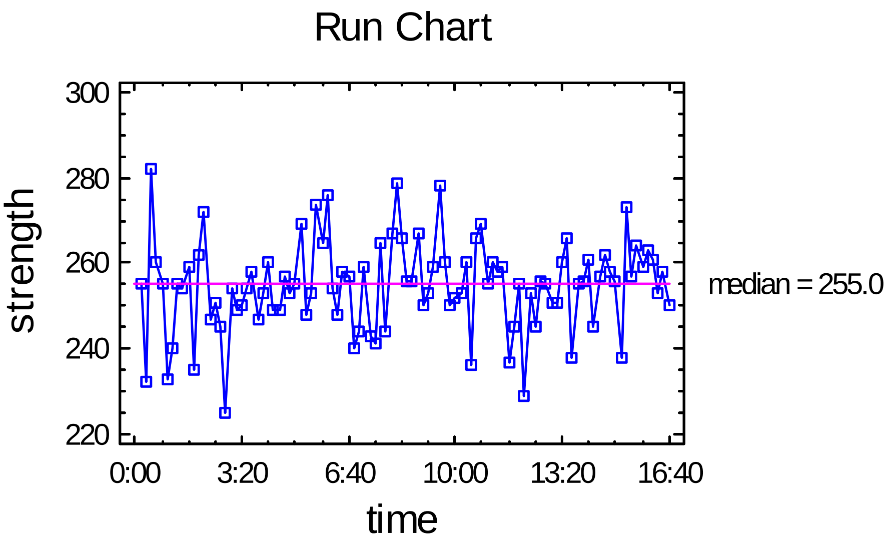
- **Subgroup Numbers or Size**: If each set of *m* rows represents a group, enter the single value *m*. For example, entering a 5 as in the example above implies that the data in rows 1-5 form the first group, rows 6-10 form the second group, and so on. If the subgroup sizes are not equal, enter the name of an additional numeric or non-numeric column containing group identifiers. The program will scan this column and place sequential rows with identical codes into the same group.

- **Date/Time/Labels:** optional column containing row identifiers, used to scale the X axis. If this field is left blank, then the groups will be numbered using integers starting at 1.

- **Select:** subset selection.

Run Chart

If the data consist of individual measurements, this pane plots the measurements in row order.

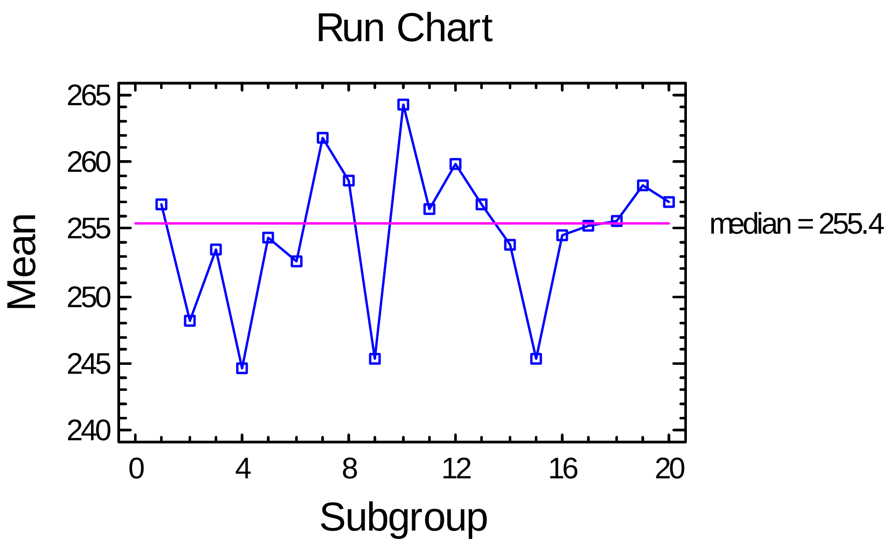


A horizontal line is drawn at the median of the values in the data column.

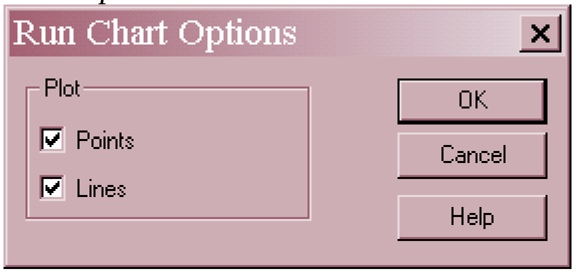
If the data is grouped, then separate run charts are created for four statistics:

1. group means
2. group medians
3. group ranges
4. group standard deviations

A chart of group means is shown below:



Pane Options



Run Chart Options

Plot

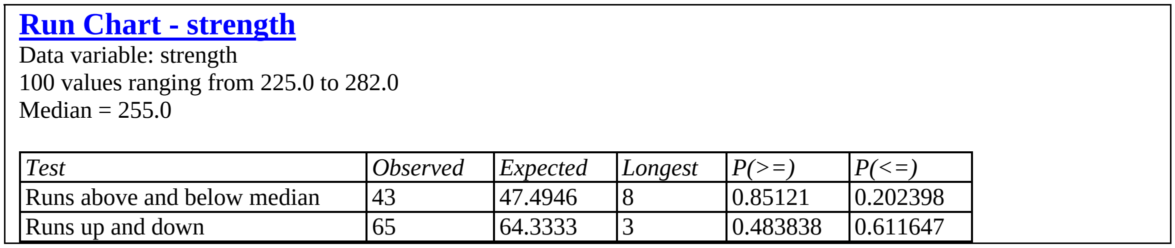
☑ Points

☑ Lines

- **Points**: plot point symbols.
- **Lines**: connect the points with a line.

Analysis Summary

The *Analysis Summary* summarizes the input data and shows the results of two tests run to determine whether or not the data plotted on each run chart can be considered a random sample from a single distribution.



Run Chart - strength

Data variable: strength

100 values ranging from 225.0 to 282.0

Median = 255.0

Test	*Observed*	*Expected*	*Longest*	*P(>=)*	*P(<=)*
Runs above and below median	43	47.4946	8	0.85121	0.202398
Runs up and down	65	64.3333	3	0.483838	0.611647

The results are important since they provide an initial test of whether or not the data come from a stable process.

The table displays the following information:

- **Observed**: the observed number of runs *k*.
- **Expected**: the expected number of runs *E(runs).*
- **Longest**: the longest run observed on the chart.
- **P(>=):** the probability of observing *k* or more runs if the data were random.
- **P(<=):** the probability of observing *k* or less runs if the data were random.

If *P(>=)* is small for the test of runs above and below the median, then there is significance evidence of **mixing** in the data, since the process crosses the median line more often than expected. If *P(<=)* is small, then there is significance evidence of **clustering** in the data, since the process does not cross the median line as often as expected.

If *P(>=)* is small for the test of runs up and down, then there is significance evidence of **oscillation** in the data, since the process changes direction more often than expected. If *P(<=)* is small, then there is significance evidence of **trending** in the data, since the process does not change direction as often as expected.

If any P value is less than 0.025, we can reject the hypothesis that the data are random samples at the 5% significance level.

Mathematical details of how the tests are conducted follow.

Test #1: runs above and below the median

This test is based on the number of times the sequence of data values remains above or below the median. A run is defined as a sequence of consecutive values all above or all below the median. (Note: values exactly equal to the median are ignored when counting runs). The test is performed as follows:

1. Calculate n_1 and n_2, the number of observations above and below the median, respectively. The sum of these two values is

$$N = n_1 + n_2 \tag{1}$$

2. Calculate k, the number of runs above and below the median.

3. Calculate the expected number of runs above and below the median if the data were a random series:

$$E(runs) = 1 + \frac{2n_1 n_2}{N} \tag{2}$$

4. Calculate the variance of the number of runs above and below the median if the data were a random series:

$$V(runs) = \frac{2n_1 n_2 (2n_1 n_2 - N)}{N^2 (N-1)} \tag{3}$$

5. Calculate the probability of observing at least k runs:

$$P(>=) = 1 - \Phi\left(\frac{k - 0.5 - E(runs)}{\sqrt{V(runs)}}\right) \tag{4}$$

where Φ(z) is the standard normal cumulative distribution function.

6. Calculate the probability of observing less than or equal to k runs:

$$P(<=) = \Phi\left(\frac{k + 0.5 - E(runs)}{\sqrt{V(runs)}}\right) \tag{5}$$

Test #2: runs up and down

This test is based on the number of times the sequence of data values either rises or falls. A run is defined as a sequence of consecutive values all going up or all going down. (Note: data values exactly equal to the previous value are ignored when counting runs). The test is performed as follows:

1. Calculate n_1 and n_2, the number of observations greater than or less than the previous value, respectively. The sum of these two values is

$$N = n_1 + n_2 \tag{6}$$

2. Calculate k, the number of runs up and down.

3. Calculate the expected number of runs up and down if the data were a random series:

$$E(runs) = \frac{2N-1}{3} \tag{7}$$

4. Calculate the variance of the number of runs up and down if the data were a random series:

$$V(runs) = \frac{16N-29}{90} \tag{8}$$

5. Calculate the probability of observing at least k runs:

$$P(>=) = 1 - \Phi\left(\frac{k - 0.5 - E(runs)}{\sqrt{V(runs)}}\right) \tag{9}$$

   where Φ(z) is the standard normal cumulative distribution function.

6. Calculate the probability of observing less than or equal to k runs:

$$P(<=) = \Phi\left(\frac{k + 0.5 - E(runs)}{\sqrt{V(runs)}}\right) \tag{10}$$